PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G10L 3/00

A1

(11) International Publication Number: WO 99/50824

(43) International Publication Date: 7 October 1999 (07.10.99)

(21) International Application Number: PCT/CA99/00258

(22) International Filing Date: 25 March 1999 (25.03.99)

(30) Priority Data:
2,230,188 27 March 1998 (27.03.98) CA

(71) Applicant *(for all designated States except US)*: HER MAJESTY THE QUEEN IN RIGHT OF CANADA as repre sented by THE MINISTER OF INDUSTRY THROUGH THE COMMUNICATION RESEARCH CENTRE [CA/CA]; 3701 Carling Avenue, P.O. Box 11490, Station H, Ottawa, Ontario K2H 8S2 (CA).

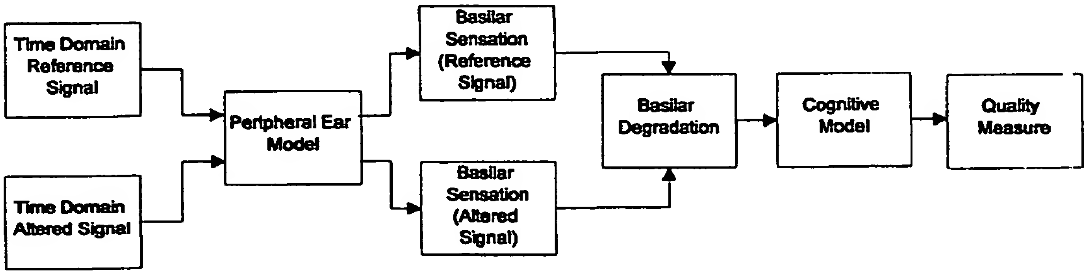
(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: TREURNIET, William, C. [CA/CA]; 25 Long Gate Court, Nepean, Ontario K2J 4E7 (CA). THIBAULT, Louis [CA/CA]; 52 Des Erables, Hull, Quebec J8Y 6K8 (CA). SOULODRE, Gilbert, Arthur, Joseph [CA/CA]; 23 Vermeer Way, Kanata, Ontario K2K 2M1 (CA).

(74) Agents: HICKS, Andrew, R. et al.; Scott & Aylen, 1000–60 Queen Street, Ottawa, Ontario K1P 5Y7 (CA).

(81) Designated States: CA, US, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(54) Title: A PROCESS AND SYSTEM FOR OBJECTIVE AUDIO QUALITY MEASUREMENT


Time Domain Reference Signal
Time Domain Altered Signal
Peripheral Ear Model
Basilar Sensation (Reference Signal)
Basilar Sensation (Altered Signal)
Basilar Degradation
Cognitive Model
Quality Measure

(57) Abstract

A process and system for providing objective quality measurement of audio signals is provided which utilizes a cognitive model for determining an objective quality measure between a reference signal and processed signal from a calculated error signal between the reference signal and processed signal.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

A Process and System for Objective Audio Quality Measurement

Field of the Invention

A process and system for providing objective quality measurement of audio signals is provided which utilizes a cognitive model for determining an objective quality measure between a reference signal and processed signal from a calculated error signal between the reference signal and processed signal.

Background of the Invention

A quality assessment of audio or speech signals may be obtained from human listeners, in which listeners are typically asked to judge the quality of a processed audio or speech sequence relative to an original unprocessed version of the same sequence. While such a process can provide a reasonable assessment of audio quality, the process is labour-intensive, time-consuming and limited to the subjective interpretation of the listeners. Accordingly, the usefulness of human listeners for determining audio quality is limited in view of these restraints. Thus, the application of audio quality measurement has not been applied to areas where such information would be useful.

For example, a system for providing objective audio quality measurement would be useful in a variety of applications where an objective assessment of the audio quality can be obtained quickly and efficiently without involving human testers each time an assessment is required. Such applications may include:

1. the assessment or characterization of implementations of audio processing equipment;

2. the evaluation of equipment or a circuit prior to placing it into service (perceptual quality line up);

3. in on-line monitoring processes to monitor audio transmissions in service;

4. in audio codec development by comparing competing encoding/compression algorithms;

5. in network planning to optimize the cost and performance of a transmission network under given constraints; and,

6. as an aid to subjective assessment (for example, as a tool for screening critical material to include in a listening test).

Current objective measures of audio or speech quality include THD (Total Harmonic Distortion) and SNR (Signal-to-Noise Ratio). The latter metric can be measured on either the time domain signal or a frequency domain representation of the signal. However, these measures are known to provide a very crude measure of audio or speech quality and are not well correlated with the subjective quality of a processed sound as compared to a test sound as determined by a human listener. Furthermore, this lack of correlation worsens when these metrics are used to measure the quality of devices such as A/D and D/A converters and perceptual audio (or speech) codecs which make use of the masking properties of the human auditory system often resulting in audio (or speech) signals being perceived as being of good or excellent quality even though the measured SNR may be poor.

Accordingly, there has been a need for an efficient system and methodology for obtaining an estimate of the perceptual quality of an audio or speech sequence that has been processed in some way. This permits frequent and automated monitoring of audio or speech equipment performance and the degree of communication network degradation. Although others have developed systems for measurement of objective perceptual quality of wide-band audio [8] [9] [10] [11], these employ algorithms that were shown to result in inadequate levels of performance in tests conducted by the ITU-R (International Telecommunications Union- Radio Communications) in 1995-1996.

Summary of the Invention

In accordance with the invention, a system is provided which enables an objective assessment of the subjective quality of a processed audio sequence relative to an original unprocessed version of the same sequence. The system assumes that both versions are simultaneously available in computer files and that they are synchronised in time. The audio sequences are processed by a computational model of hearing which removes auditory components from the input that are normally not perceptible by human listeners. The result is a numerical representation of the pattern of excitation produced by the sounds on the basilar membrane of the human auditory system. The basilar sensation level of the processed version is compared with that of the unprocessed version, and the difference is used to predict the average quality rating that would be expected from human listeners.

In accordance with the invention, there is provided a process for determining an objective audio quality measurement of a processed audio sequence relative to a corresponding unprocessed audio sequence, comprising the steps of:

a) passing the unprocessed audio sequence and the processed audio sequence through an auditory model to create a basilar degradation signal of the unprocessed and processed audio sequences;

b) calculating at least one input variable from the basilar degradation signal, the at least one input variable selected from any one of or a combination of average distortion level, maximum distortion level, average reference level, reference level at maximum distortion, coefficient of variation of distortion, and correlation between reference and distortion patterns;

c) calculating a harmonic structure in the distortion variable from an error spectrum obtained through a comparison of the unprocessed and processed audio sequences; and,

d) passing the at least one input variable from step b) and the harmonic structure in the distortion variable from step c) through a cognitive model utilizing a multi-layer neural network to obtain an objective quality measure of the processed audio sequence with respect to the unprocessed audio sequence.

In a further embodiment, the number of input variables selected in step b) is determined by the desired accuracy of the quality measure.

Preferably, step b) includes calculating the basilar degradation signal using any one of or a combination of a level-dependent or frequency dependent spreading function having a recursive filter, and/or includes calculating the basilar degradation signal using a recursive filter implementation of a spreading function.

Still further, and preferably, step d) includes calculating separate weightings for adjacent frequency ranges for use in the cognitive model and the basilar degradation signal is used to calculate any one of or a combination of perceptual inertia, perceptual asymmetry and adaptive threshold for rejection of relatively low values for use within the cognitive model.

In a further embodiment, there is provided a system for determining an objective audio quality measurement of an unprocessed audio sequence and a corresponding processed audio sequence comprising:

an auditory model module for providing a basilar degradation signal of the unprocessed and processed audio sequences;

a first variable processing module for calculating at least one input variable from the basilar degradation signal, the first variable processing module for calculating at least one input variable selected from any one of or a combination of average distortion level, maximum distortion level, average reference level, reference level at maximum distortion, coefficient of variation of distortion, and correlation between reference and distortion patterns;

a second variable processing module for calculating a harmonic structure in the distortion variable from an error spectrum obtained through a comparison of the unprocessed and processed audio sequences;

a cognitive model module for receiving the at least one input variable from the first variable processing module and the harmonic structure in the distortion variable from the second variable processing module, the cognitive model module utilizing a mulit-layer neural network to obtain an objective quality measure of the processed audio sequence with respect to the unprocessed sequence from the at least one input variable and harmonic structure in the distortion variable.

Alternate embodiments of the invention include an algorithm for calculating the basilar degradation signal using any one of or a combination of a level-dependent or frequency dependent spreading function having a recursive filter, calculating the basilar degradation signal using a recursive filter implementation of a spreading function, calculating separate weightings for adjacent frequency ranges, and/or calculating any one of or a combination of perceptual inertia, perceptual asymmetry and adaptive threshold for rejection of relatively low values for use within the cognitive model from the basilar degradation signal.

The system may also include input means for introducing the processed and unprocessed audio sequences into the system.

Brief Description of the Drawings

In the accompanying drawings:

Figure 1 is a high level representation of a peripheral ear and cognitive model of audition developed as a tool for objective evaluation of the perceptual quality of audio signals;

Figure 2A shows successive stages of processing of the peripheral ear model;

Figure 2B shows a flow chart of the processing of a reference and test signal to obtain a quality measurement;

Figure 3 shows a representative reference power spectrum;

Figure 4 shows a representative test power spectrum;

Figure 5 shows a representative middle ear attenuation spectrum of the reference signal;

Figure 6 shows a representative middle ear attenuation spectrum of the test signal;

Figure 7 shows a representative error spectrum from the reference and test signals;

Figure 8 shows a representative error cepstrum from the reference and test signals;

Figure 9 shows a representative excitation spectrum from the reference signal;

Figure 10 shows a representative excitation spectrum from the test signal;

Figure 11 shows a representative excitation error signal; and,

Figure 12 shows a representative echoic memory output signal.

Detailed Description of the Invention

1- Overview of the Auditory Process of the Human Ear

In developing a system for objective audio quality measurement, consideration of the physical shape and performance of the ear is required. The primary regions of the ear include an outer portion, a middle portion and an inner portion. The outer ear is a partial barrier to external sounds and attenuates the sound as a function of frequency. The ear drum, at the end of the ear canal, transmits the sound vibrations to a set of small bones in the middle ear. These bones propagate the energy to the inner ear via a small window in the cochlea. A spiral tube within the cochlea contains the basilar membrane that resonates to the input energy according to the frequencies present. That is, the location of vibration of the membrane for a given input frequency is a monotonic, non-linear function of frequency. The distribution of mechanical energy along the membrane is called the excitation pattern. The mechanical energy is transduced to neural activity via hair cells connected to the basilar membrane, and the distribution of neural activity is passed to the brain via the fibres in the auditory nerve.

2 - Model Development

2.0 Summary of Psychoacoustic (Peripheral Ear) and Cognitive Model

In order to provide an objective audio quality measurement system, simulation of the peripheral auditory processes is required to create a basilar membrane representation of an audio signal. Thereafter, in order to assess the quality of the audio sequence, the basilar membrane representation of the audio signal is subsequently subjected to simple transformations based on assumptions about higher level perceptual processing, in order to provided an estimated perceptual quality of the signal relative to a known reference signal. Calibration of the system is required using data obtained from human observers in a number of listening tests. A high level representation of the system is shown in Fig. 1.

With reference to Figure 1, an unprocessed audio signal and processed audio signal are passed through a mathematical auditory model of the human ear (peripheral ear) in which components of the signals are masked in a manner approximating the masking of a signal in the human ear. The resulting output, referred to as the basilar representation or basilar signal, from both the unprocessed and processed signals are compared to create an indication of the relative differences between the two signals, referred to as the basilar degradation signal, also referred to as the excitation error. The basilar degradation signal is essentially an error signal representing the error between the unprocessed and processed signals that has not been masked by the peripheral ear model.

The basilar degradation signal is passed to the cognitive model which, through the use of a number of variables, outputs an objective perceptual quality rating based on the monaural degradations as well as any shifts in the position of the binaural auditory image.

2.1.1 The Auditory (Peripheral Ear) Model

The auditory (peripheral ear) model is designed to model the underlying physical phenomena of simultaneous masking effects within the ear. That is, the model considers the transfer characteristics of the middle and inner ear to form a representation of the signal corresponding to the mechanical to neural processing of the middle and inner ear . The model assumes that:

1) the mechanical phenomena of the inner ear are linear but not necessarily invariant with respect to amplitude and frequency. That is, the spread of energy in the inner ear may, if desired, be made a function of signal amplitude and frequency.
2) the basilar membrane is sensitive to input energy according to a logarithmic sensitivity function.
3) the basilar membrane has poor temporal resolution.

The input signals (both the unprocessed and processed signals) are processed as follows:

1. The input signal 21 is decomposed into a time-frequency representation, to an energy spectrum 23 using a discrete fournier transform (DFT) 22. Typically, a Hann window of approximately 40 msec is applied to the input data, with a 50 percent

overlap between successive windows.

2. The energy spectrum 23 is multiplied by a frequency dependent function 24 which models the effect of the ear canal and the middle ear to produce an attenuated energy spectrum 25.

3. The attenuated spectral energy values 25 are mapped 26 from the frequency scale to a pitch scale to create a localized basilar energy representation 27 that is more linear with respect to both the physical properties of the inner ear and observed psychophysical effects.

4. The localized basilar energy representations 27 are then convolved with a spreading function to simulate the dispersion of energy along the basilar membrane to create a dispersed energy representation 29.

5. The dispersed energy representation 29 is adjusted through the addition of an intrinsic frequency-dependent energy to each pitch component to account for the absolute threshold of hearing.

6. Conversion of the energy to decibels results in a basilar membrane representation (sensation) 31 of the signal.

These successive stages of processing are shown in Fig. 2.

More specifically, the following computations are performed for each block:

Box 24

The energy spectrum 23 is multiplied by an attenuation spectrum of a low pass filter which models the effect of the ear canal and the middle ear. The attenuation spectrum, described by the following equation, is modified from that presented in reference [3] in order to extend the high frequency cutoff. This was accomplished by changing the exponent in equation 1 from 4.0 to 3.6.

(equation 1) $A_{dB} = -6.5\, e^{(-0.6(f-0.33)^2)} + 10^{-3} f^{3.6}$

where A is the attenuated value in decibels.

Box 26

The attenuated spectral energy values 25 are transformed using a non-linear mapping function from the frequency domain to the subjective pitch domain using the bark scale (an equal interval pitch scale). A commonly used mapping function [5] is as follows:

(equation 2) $B = 1300\arctan(0.76f) + \arctan(f/7.5)^2$

where B is pitch on the Bark scale.

A new function was created that improves resolution at higher frequencies. A simpler expression was derived for this new function, and is presented as follows (the frequency f is in Hz.):

(equation 3) $p = f/(9.0304615\mathrm{e}-05 f + 2.6167612)$

where p is pitch.

Box 28

The basilar membrane components are convolved with a spreading function to simulate the dispersion of energy along the basilar membrane. The spreading function applied to a pure tone results in an asymmetric triangular excitation pattern with slopes that may be selected to optimize performance. The spreading is implemented by sequentially applying two IIR filters,

(equation 4 and 5) $H_1(z) = 1/(1-a/z)$ and $H_2(z) = 1/(1-bz)$

where the a and b coefficients are the reciprocals of the slopes of the spreading function on the dB scale.

Box 30

A spreading function with a slope on the low frequency side (LSlope) of 27 dB/Bark and a slope on the high frequency side of -10 dB/Bark has been implemented. For the frequency-to-pitch mapping function given above (equation 3), it has been found that predictions of audio quality ratings improved with fixed spreading function slopes of 24 and -4 dB/Bark, respectively.

Other comments with respect to Spreading Functions

The literature (e.g. [3]) indicates that the slope *S* of the spreading function on the low frequency side should be fixed (i.e., *Lslope* = 0.27 dB/mel). However, on the high frequency side, the slope is said to vary with both signal level and frequency. That is, it increases with both decreasing level and increasing frequency. The following equation for computing the higher frequency slope (S) as a function of frequency and level is based on an equation in [3]:

(equation 6) $$S_{dB/mel} = HSlope - LRate \cdot L_{dB} + FRate/F_{KHz}$$

Suggested values of 24 for *HSlope*, 230 for *FRate* and 0.2 for L*Rate* . However, in a peripheral ear model, the best values for these parameters are dependent on other system components such as the frequency to pitch mapping function.

In the subject system, parameter values for a particular system configuration using a function optimization procedure have been made. Optimal values are those that minimize the difference between the model's performance and a human listener's performance in a signal detection experiment. This procedure allows the model parameters to be tailored so that it behaves like a particular listener - reference [6].

In other psychoacoustic models, the spreading function is applied to each pitch position by distributing the energy to adjacent positions according to the magnitude of the spreading function at those positions. Then the respective contributions at each position are added to obtain the total energy at that position. Dependence of the spreading function slope on level and frequency is accommodated by dynamically selecting the slope that is appropriate for the instantaneous level and frequency.

In the subject system, to implement the dependence of the slope on level using the IIR filter implementation, a new procedure was developed. It is important to note that because convolution is a linear operation, the effects of convolving data with different spreading functions may be summed. Therefore, input values within particular ranges are convolved with level-specific spreading functions, and the results summed to approximate a single convolution with the desired dependence on signal level. Accuracy of the result may be traded off with computational load by varying the number of signal quantization levels.

A similar procedure may be used to include the dependence of the slope on both level and frequency. That is, the frequency range may also be divided into subranges, and levels within each subrange are convolved with the level and frequency-specific IIR filters.

Again, the results are summed to approximate a single convolution with the desired dependence on signal level and frequency.

2.2 Cognitive Model

Since the basilar membrane representation produced by the peripheral ear model is expected to represent only supraliminal aspects of the input audio signal, this information is the basis for simulating results of listening experiments. That is, ideally, the basilar sensation vector produced by the auditory model represents only those aspects of the audio signal that are perceptually relevant.

However, the perceptual salience of audible basilar degradations can vary depending on a number of contextual or environmental factors. Therefore, the reference basilar membrane representation (ie the unprocessed basilar representation) and the basilar degradation vectors (ie the basilar degradation signal) are processed in various ways according to reasonable assumptions about human cognitive processing.

The result of processing according to the cognitive model is a number of variables, described below, that singly or in combination produce a perceptual quality rating. While other methods also calculate a quality measurement using one or more variables derived from a basilar membrane representation (e.g., [11][12]), these methods use different variables and combinations of variables to produce an objective quality measurement. The use of these variables is novel and have not been used previously to measure audio quality.

Preferably, the peripheral ear model processes a frame of data every 21 msec. Calculations for each frame of data are reduced to a single number at the end of a 20 or 30 second audio sequence.

The most significant variables are:

1. average distortion level;
2. maximum distortion level;
3. average reference level;
4. reference level at maximum distortion;
5. coefficient of variation of distortion;

6. correlation between reference and distortion patterns; and,

7. harmonic structure in the distortion.

In the cognitive model, a value for each variable is computed for each of a discrete number of adjacent frequency ranges. This allows the values for each range to be weighted independently, and also allows interactions among the ranges to be weighted. Three ranges are usually employed - 0 to 1000 Hz, 1000 to 5000 Hz, and 5000 to 18000 Hz. An exception is the measure of harmonic structure of spectrum error that is calculated using the entire audible range of frequencies.

Accordingly,18 variables result from the first six variables listed above when the three pitch ranges are considered in addition to the harmonic structure in the distortion variable for a total of 19 variables. The variables are mapped to a mean quality rating of that audio sequence as measured in listening tests using a multi-layer neural network. Non-linear interactions among the variables are required because the average and maximum errors should be weighted differentially as a function of the coefficient of variation. The use of a multilayer neural network with semi-linear activation functions allows this possibility.

The feature calculations and the mapping process implemented by the neural network constitute a task-specific model of auditory cognition.

2.2.1 Cognitive Model Pre-Processing Calculations

Prior to processing within the cognitive model, a number of pre-processing calculations are performed as described below. Essentially, these pre-processing calculations are performed in order to address the fact that the perceptability of distortions is likely affected by the characteristics of the current distortion as well as temporally adjacent distortions. Thus, the pre-processing considers:

Perceptual Inertia

A particular distortion is considered inaudible if it is not consistent with the immediate context provided by preceding distortions. This effect is herein defined as perceptual inertia. That is, if the sign of the current error is opposite to the sign of the average error over a short time interval, the error is considered inaudible. The duration of this memory is close to 80 msec, which is the approximate time for the asymptotic integration of loudness of a constant energy stimulus by human listeners – reference [6].

In practice, the energy is accumulated over time, and data from several successive frames determine the state of the memory. At each time step, the window is shifted one frame and each basilar degradation component is summed algebraically over the duration of the window. Clearly, the magnitudes of the window sums depend on the size of the distortions, and whether their signs change within the window. The signs of the sums indicate the state of the memory at that extended instant in time.

The content of the memory is updated with the distortions obtained from processing the current frame. However, the distortion that is output at each time step is the rectified input, modified according to the relation of the input to the signs of the window sums. If the input distortion is positive and the same sign as the window sum, the output is the same as the input. If the sign is different, the corresponding output is set to zero since the input does not continue the trend in the memory at that position. In particular, the output distortion at the *i*th position, D_i, is assigned a value depending on the sign of the *i*th window mean, W_i and the *i*th input distortion, E_i.

```
If ( SGN( $E_i$ ) EQ SGN( $W_i$ ) AND $E_i$ GT 0.0 )$D_i$ =  $E_i$
If ( SGN( $E_i$ ) NE SGN( $W_i$ ) )                 $D_i$ = 0.0
```

Perceptual Asymmetry

Negative distortions are treated somewhat differently. There are indications in the literature on perception - references [2][4] - that information added to a visual or auditory display is more readily identified than information taken away. Accordingly, this program weighs less heavily the relatively small distortions resulting from spectral energy removed from, rather than added to, the signal being processed. Because it is considered less noticeable, a small negative distortion receives less weight than a positive distortion of the same magnitude. As the magnitude of the error increases, however, the importance of the sign of the error should decrease. The size of the error at which the weight approaches unity was somewhat arbitrarily chosen to be *Pi*, as shown in the following equation.

```
If ( SGN( $E_i$ ) EQ SGN( $W_i$ ) AND $E_i$ LT 0.0 )
```

$$D_i = |E_i| * \arctan(\ 0.5 * |E_i|)$$

where | | represents the absolute value and * is the scalar multiplication.

Adaptive Threshold for Averaging

The distortion values obtained from the memory could be reduced to a scalar simply by averaging. However, if some pitch positions contain negligible values, the impact of significant adjacent narrow band distortions would be reduced. Such biasing of the average could be prevented by ignoring all values under a fixed threshold, but frames with all distortions under that threshold would then have an average distortion of zero.

This also seems like an unsatisfactory bias. Instead, an adaptive threshold has been chosen for ignoring relatively small values. That is, distortions in a particular pitch range are ignored if they are less than a fraction (eg. one-tenth) of the maximum in that range.

The average distortion over time for each pitch range is obtained by summing the mean distortion across successive non-zero frames. A frame is classified as non-zero when the sum of the squares of the most recent 1024 input samples exceeds 8000 (i.e., more than 9 dB per sample on average).

2.2.2 Cognitive Model Variables

2.2.2.1 Average Distortion Level

For each analysis frame, the perceptual inertia and perceptual assymetry characteristics of the cognitive model transforms the basilar error vector into an echoic memory vector which describes the extent of degradation over the entire range of auditory frequencies. These resulting values are averaged for each pitch range with the adaptive threshold set at 0.1 of the maximum value in the range, and the final value is obtained by a simple average over the frames.

2.2.2.2 Maximum Distortion Level

The maximum distortion level is obtained for each pitch range by finding the frame with the maximum distortion in that range. The maximum value is emphasized for this calculation by defining the adaptive threshold as one-half of the maximum value in the given pitch range instead of one-tenth that is used above to calculate the average distortion.

2.2.2.3 Average Reference Level

The average reference level over time is obtained by averaging the mean level of the reference signal in each pitch range across successive non-zero frames.

2.2.2.4 Reference Level at Maximum Distortion

The value of this variable in each pitch region is the reference level that corresponds to the maximum distortion level calculated as described above.

2.2.2.5 Coefficient of Variation of Distortion

The coefficient of variation is a descriptive statistic that is defined as the ratio of the standard deviation to the mean [10]. The coefficient of variation of the distortion over frames has a relatively large value when a brief, loud distortion occurs in an audio sequence that otherwise has a small average distortion. In this case, the standard deviation is large compared to the mean. Since listeners tend to base their quality judgments on this brief but loud event rather than the overall distortion, the coefficient of variation may be used to differentially weight the average distortion versus the maximum distortion in the audio sequence. It is calculated independently for each pitch region.

2.2.2.6 Correlation of reference and distortion spectra

When the peak magnitudes of the distortion coincide in pitch with the peak magnitudes of the reference signal, perceptibility of the distortion may be differentially affected. The correlation C between the distortion (E) and reference (R) vectors should reflect this coincidence, and this is found by calculating the cosine of the angle between the vectors for each pitch region as follows:

$$C = \frac{\vec{R} \bullet \vec{E}}{|\vec{R}| * |\vec{E}|}$$

where •is the dot product operator, | |is the magnitude of the enclosed vector and * is the scalar multiplication.

2.2.2.7 Harmonic structure in distortion

Listeners may respond to some structure of the error within a frame, as well as to its magnitude. Harmonic structure in the error can result, for example, when the reference signal has strong harmonic structure, and the signal under test includes additional broadband noise. In that case, masking is more likely to be inadequate at frequencies where the level of the reference signal is low between the peaks of the harmonics. The result would be a periodic structure in the error that corresponds to the structure in the original signal.

The harmonic structure is measured in either of two ways. In the first method, it is described by the location and magnitude of the largest peak in the spectrum of the log energy autocorrelation function. The correlation is calculated as the cosine between two vectors.

In the second method, the periodicity and magnitude of the harmonic structure is inferred from the location of the peak with the largest value in the cepstrum of the error. The relevant parameter is the magnitude of the largest peak. In some cases, it is useful to set the magnitude to zero if the periodicity of the error is significantly different from that of the reference signal. Specifically, if the difference between the two periods is greater than one-quarter of the reference period, the error is assumed to have no harmonic structure related to the original signal.

2.2.3 Calibration of Model Predictors

The mean quality ratings obtained from human listening experiments is predicted by a weighted non-linear combination of the 19 variables described above. The prediction algorithm was optimised using a multilayer neural network to derive the appropriate weightings of the input variables. This method permits non-linear interactions among the variables which is required to differentially weight the average distortion and the maximum distortion as a function of the coefficient of variation.

The system relating the above variables to human quality ratings was calibrated using data from eight different listening tests that used the same basic methodology. These experiments were known in the ITU-R Task Group 10/4 as MPEG90, MPEG91, ITU92CO, ITU92DI, ITU93, MPEG95, EIA95, and DB2. Generalization testing was performed using data from the DB3 and CRC97 listening tests.

3. Examples

With reference to Figures 2B-12, examples of the processing of a representative reference signal and test signal is described. Figures 3-4 show a typical reference spectrum (box 100 and Figure 3) and test spectra (box 102 Figure 4).

These spectra are processed by the peripheral ear model (boxes 104 and 106, Figure 5 and 6) to provide representative masking by the outer and middle ear. The basilar representation or excitation (boxes 108 and 110) are shown in Figures 9 and 10 and subsequently compared (box 111) to provide an excitation error signal (box 112) as shown in Figure 11. Pre-processing of the excitation error signal (box 114) as shown in Figure 12 determines the effects of perceptual inertia and asymmetry for use within the cognitive model (box 116).

Additional input for the cognitive model is provided by a comparison 118 of the reference and test spectra (boxes 100 and 102) to create an error spectrum (box 120) as shown in Figure 7 The error spectrum (box 120) is used to determine the harmonic structure (box 122, Figure 8) for use within the cognitive model (box 116). The cognitive model provides a discrete output of the objective quality of the test signal through the calculation, averaging and weighting of the input variables through a multi-layer neural network.

The number of cognitive model variables utilized to provide an objective quality measure is dependent on the desired level of accuracy in the quality measure. That is, an increased level of accuracy will utilize a larger number of cognitive model variables to provide the quality measure. Experimentally, it has been found that a combination of the above variables provides the best objective measurement of audio quality.

4. Implementation

The system and process of the invention are implemented using appropriate computer systems enabling the processed and unprocessed audio sequences to be collected and processed. Appropriate computer processing modules are utilized to process data within the peripheral ear model and cognitive model in order to provide the desired objective quality measure. The system may also include appropriate hardware inputs to allow the input of processed and unprocessed audio sequences into the system.

References

[1] M. Florentine and S. Buus. An excitation-pattern model for intensity discrimination. *J. Acoust. Soc. Am.*, 70:1646-1654, 1981.

[2] E. Hearst. Psychology and nothing. *American Scientist*, 79:432-443, 1979.

[3] E. Terhardt, G. Stoll, M. Sweeman. Algorithm for extraction of pitch and pitch salience from complex tonal signals. *J. Acoust. Soc. Am.* 71(3):678-688, 1982.

[4] M. Treisman. Features and objects in visual processing. *Scientific American*, 255[5]:114-124, 1986.

[5] E. Zwicker and E. Terhardt. Analytical expressions for critical-band rate and critical bandwidth as a fuction of frequency. *J. Acoust. Soc. Am.* 68(5): 1523-1525, 1980.

[6] Treurniet, W.C. Simulation of individual listeners with an auditory model. *Proceedings of the Audio Engineering Society*, Copenhagen, Denmark, Reprint Number 4154, 1996.

[7] B. Paillard, P. Mabilleau, S. Morisette, and J. Soumagne. Perceval: Perceptual evaluation of the quality of audio signals, *J. Audio Eng. Soc.*, Vol. 40, pages 21-31, 1992.

[8] J.G. Beerends and J.A. Stemerdink, A perceptual audio quality measure based on a psychoacoustic sound representation, *J. Audio Eng. Soc.*, Vol. 40, pages 963-978, December 1992.

[9] C. Colomes, M. Lever, J.B. Rault, and Y.F. Dehery, A perceptual model applied to audio bit-rate reduction, *J.Audio Eng. Soc.*, Vol. 43, pages 233-240, April 1995.

[10] K. Brandenburg and T. Sporer. 'NMR' and 'Masking Flag': Evaluation of quality using perceptual criteria, *11th International AES Conference on Audio*

Test and Measurement, Portland, 1992, pp169-179.

[11] T. Thiede and E. Kabot, A New Perceptual Quality Measure for Bit Rate Reduced Audio *Proceedings of the Audio Engineering Society*, Copenhagen, Denmark, Reprint Number 4280, 1996.

[12] J.G. Beerends, Measuring the quality of speech and music codecs, an integrated psychoacoustic approach. *Proceedings of the Audio Engineering Society*, Copenhagen, Denmark, Reprint Number 4154, 1996.

Claims:

1. A process for determining an objective audio quality measurement of a processed audio sequence relative to a corresponding unprocessed audio sequence, comprising the steps of:

a) passing the unprocessed audio sequence and the processed audio sequence through an auditory model to create a basilar degradation signal of the unprocessed and processed audio sequences;

b) calculating at least one input variable from the basilar degradation signal, the at least one input variable selected from any one of or a combination of average distortion level, maximum distortion level, average reference level, reference level at maximum distortion, coefficient of variation of distortion, and correlation between reference and distortion patterns;

c) calculating a harmonic structure in the distortion variable from an error spectrum obtained through a comparison of the unprocessed and processed audio sequences; and,

d) passing the at least one input variable from step b) and the harmonic structure in the distortion variable from step c) through a cognitive model utilizing a multi-layer neural network to obtain an objective quality measure of the processed audio sequence with respect to the unprocessed audio sequence.

2. A process as in claim 1 wherein the number of input variables selected in step b) is determined by the desired accuracy of the quality measure.

3. A process as in any one of claims 1-2 wherein step b) includes calculating the basilar degradation signal using any one of or a combination of a level-dependent or frequency

dependent spreading function having a recursive filter.

4. A process as in any one of claims 1-3 wherein step b) includes calculating the basilar degradation signal using a recursive filter implementation of a spreading function.

5. A process as in any one of claims 1-4 wherein step d) includes calculating separate weightings for adjacent frequency ranges for use in the cognitive model.

6. A process as in any one of claims 1-5 wherein prior to step d), the basilar degradation signal is used to calculated any one of or a combination of perceptual inertia, perceptual asymmetry and adaptive threshold for rejection of relatively low values for use within the cognitive model.

7. A system for determining an objective audio quality measurement of an unprocessed audio sequence and a corresponding processed audio sequence comprising:

an auditory model module for providing a basilar degradation signal of the unprocessed and processed audio sequences;

a first variable processing module for calculating at least one input variable from the basilar degradation signal, the first variable processing module for calculating at least one input variable selected from any one of or a combination of average distortion level, maximum distortion level, average reference level, reference level at maximum distortion, coefficient of variation of distortion, and correlation between reference and distortion patterns;

a second variable processing module for calculating a harmonic structure in the distortion variable from an error spectrum obtained through a comparison of the unprocessed and processed audio sequences;

a cognitive model module for receiving the at least one input variable from the first variable processing module and the harmonic structure in the distortion variable from the second variable processing module, the cognitive model module utilizing a multi-layer neural network to obtain an objective quality measure of the processed audio sequence with respect to the unprocessed sequence from the at least one input variable and harmonic structure in the distortion variable.

8. A system as in claim 7 wherein the first variable processing module includes an algorithm for calculating the basilar degradation signal using any one of or a combination of a level-dependent or frequency dependent spreading function having a recursive filter.

9. A system as in any one of claims 7-8 wherein the first variable processing module includes calculating the basilar degradation signal using a recursive filter implementation of a spreading function.

10. A system as in any one of claims 7-9 wherein the cognitive model module includes an algorithm for calculating separate weightings for adjacent frequency ranges.

11. A system as in any one of claims 7-10 further comprising an algorithm for calculating any one of or a combination of perceptual inertia, perceptual asymmetry and adaptive threshold for rejection of relatively low values for use within the cognitive model from the basilar degradation signal.

12. A system as in any one of claims 7-11 further comprising input means for introducing the processed and unprocessed audio sequences into the system.

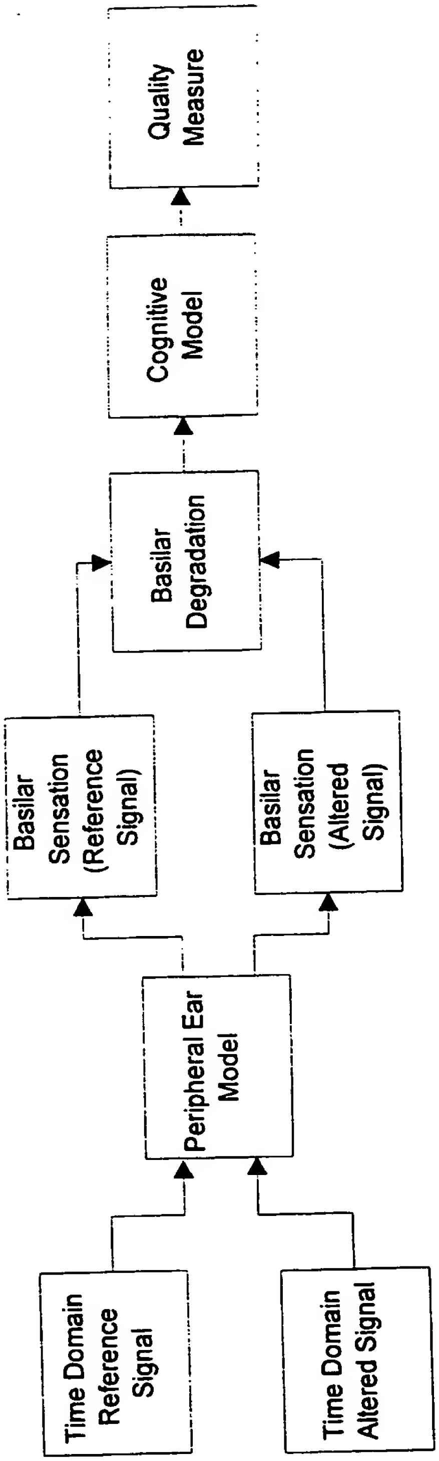
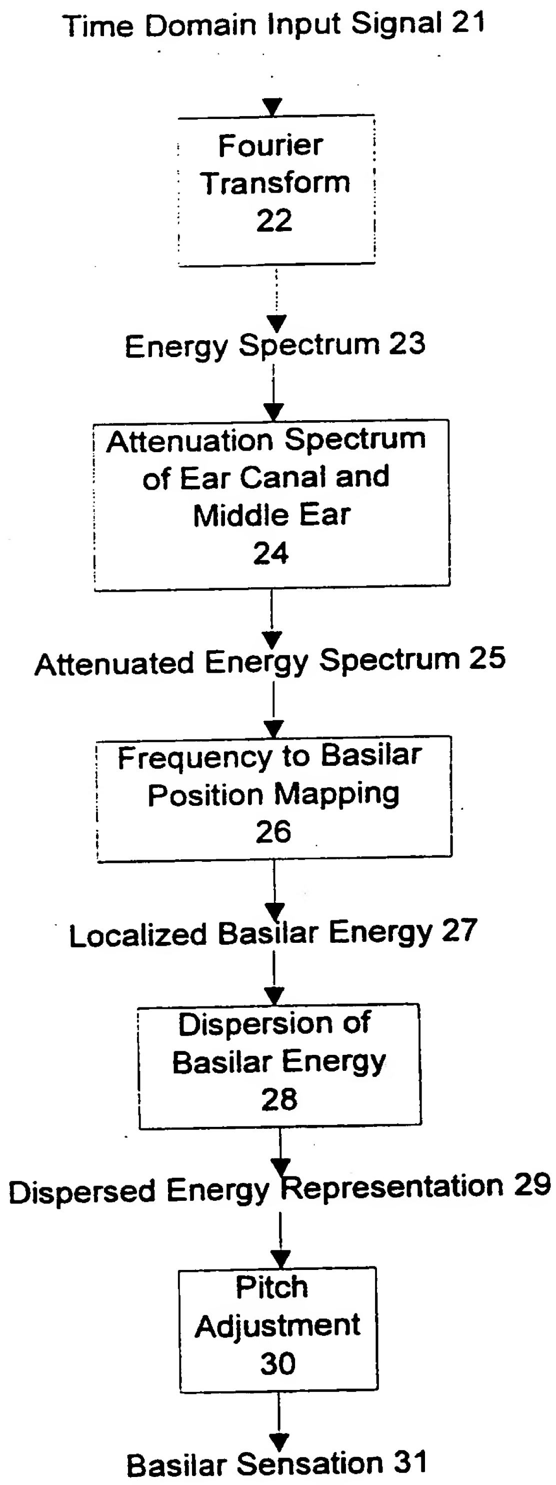

Time Domain Reference Signal
Time Domain Altered Signal
Peripheral Ear Model
Basilar Sensation (Reference Signal)
Basilar Sensation (Altered Signal)
Basilar Degradation
Cognitive Model
Quality Measure

Figure 1

Figure 2A
Auditory (Peripheral Ear) Model


Time Domain Input Signal 21
Fourier Transform 22
Energy Spectrum 23
Attenuation Spectrum of Ear Canal and Middle Ear 24
Attenuated Energy Spectrum 25
Frequency to Basilar Position Mapping 26
Localized Basilar Energy 27
Dispersion of Basilar Energy 28
Dispersed Energy Representation 29
Pitch Adjustment 30
Basilar Sensation 31

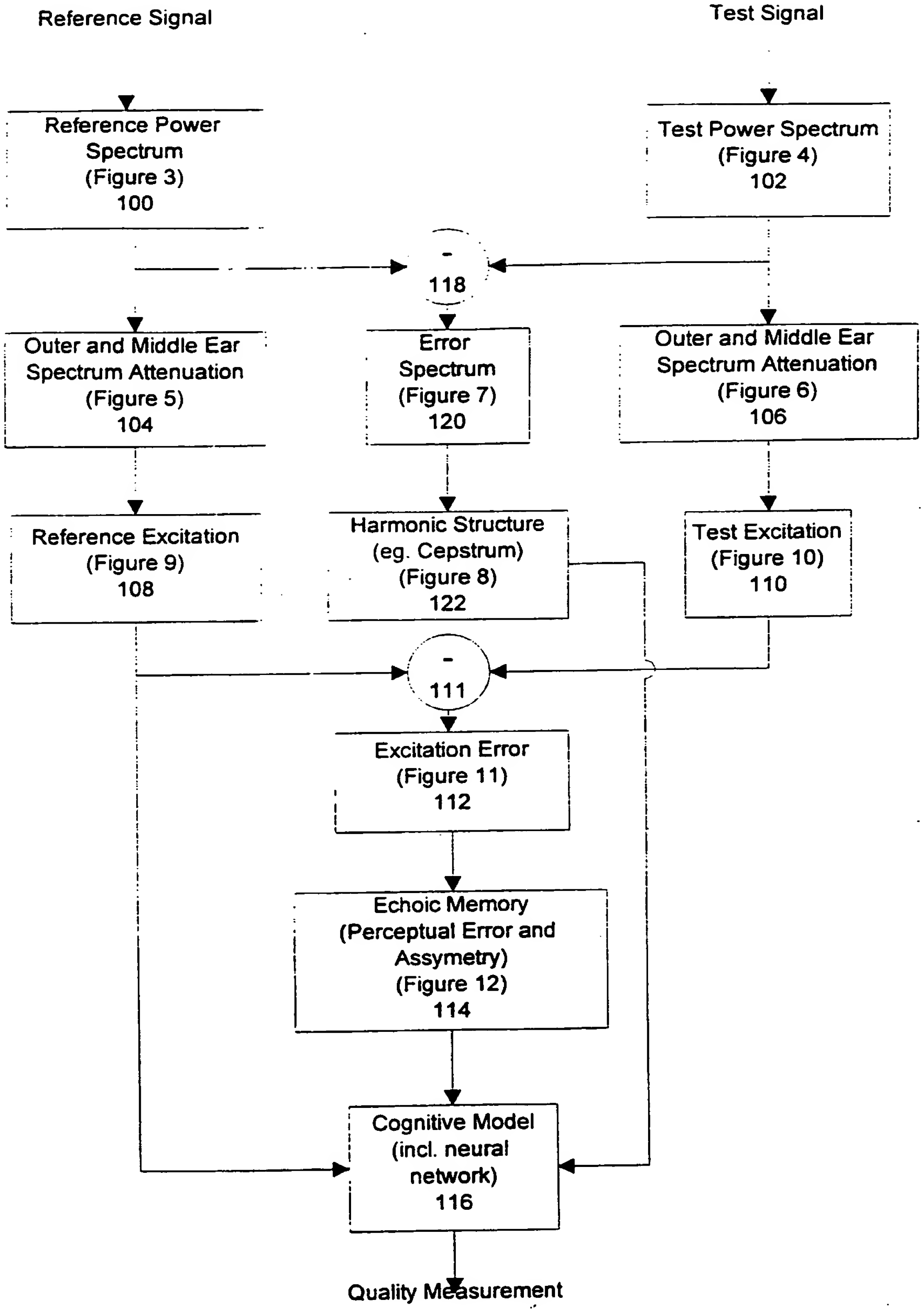

Reference Signal
Test Signal
Reference Power Spectrum (Figure 3) 100
Test Power Spectrum (Figure 4) 102
-
118
Outer and Middle Ear Spectrum Attenuation (Figure 5) 104
Error Spectrum (Figure 7) 120
Outer and Middle Ear Spectrum Attenuation (Figure 6) 106
Reference Excitation (Figure 9) 108
Harmonic Structure (eg. Cepstrum) (Figure 8) 122
Test Excitation (Figure 10) 110
-
111
Excitation Error (Figure 11) 112
Echoic Memory (Perceptual Error and Assymetry) (Figure 12) 114
Cognitive Model (incl. neural network) 116
Quality Measurement

Figure 2B Model Flow Chart

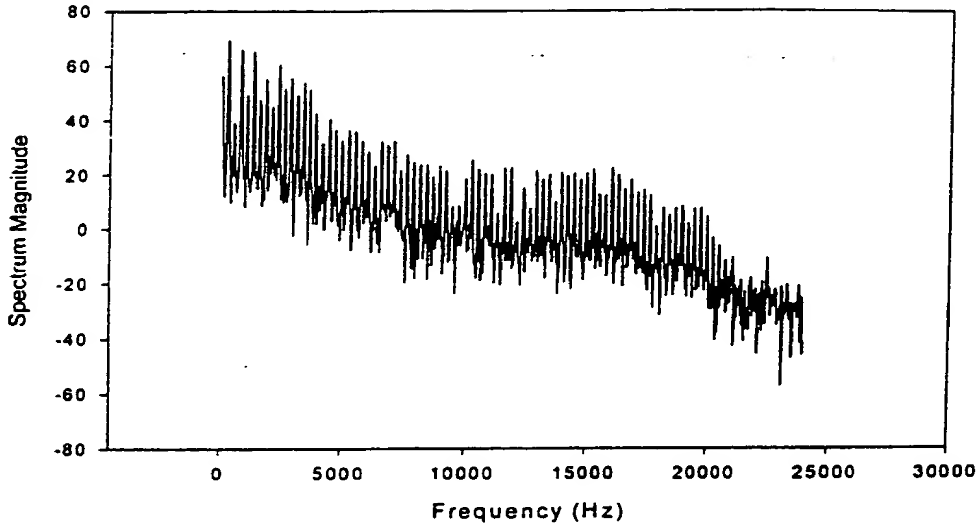


Figure 3. Reference spectrum

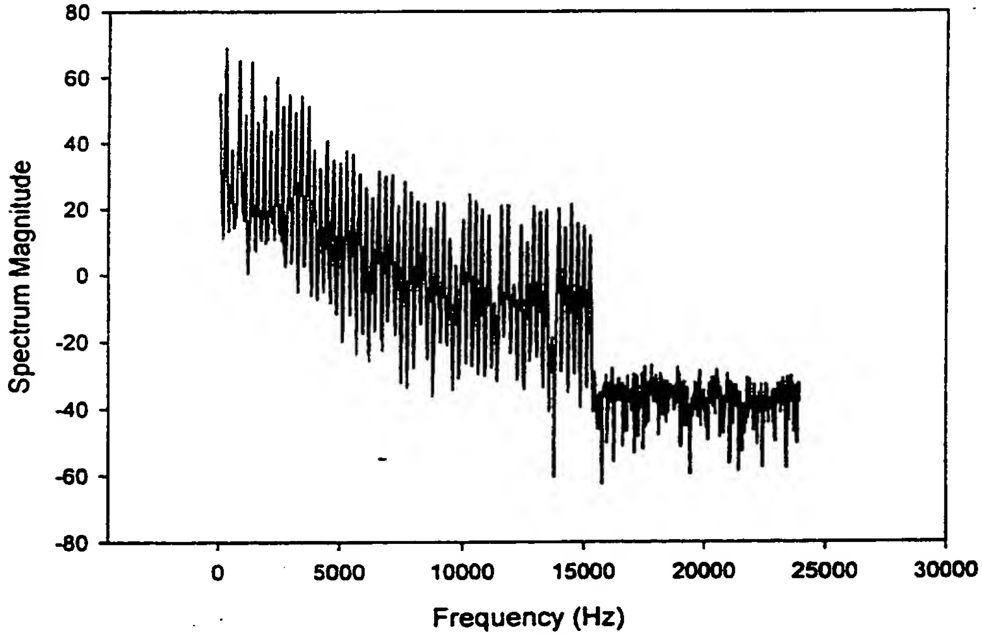


Figure 4. TestA spectrum

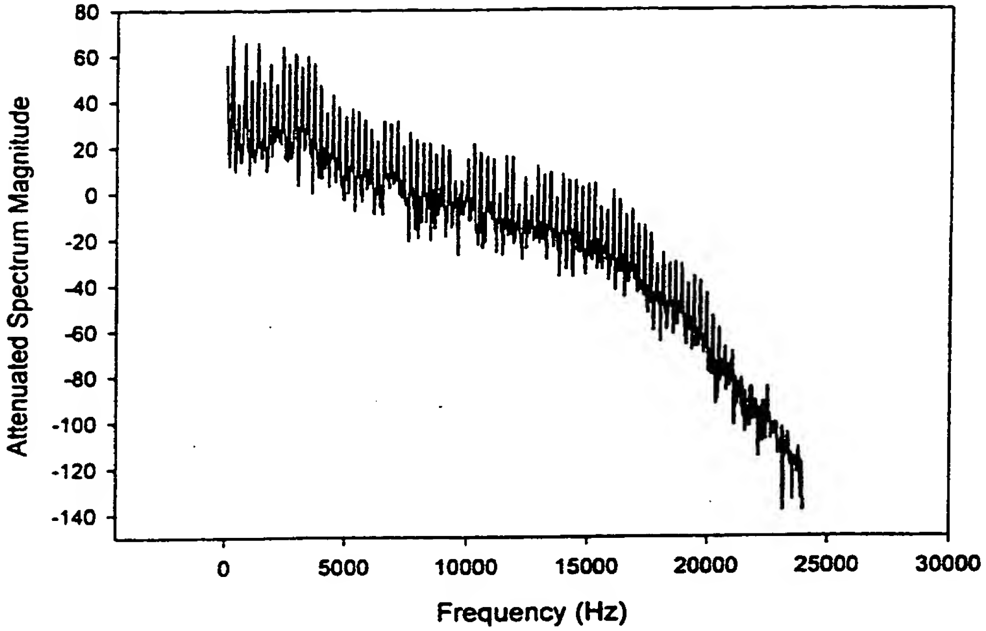


Figure 5. Middle ear attenuation of reference spectrum

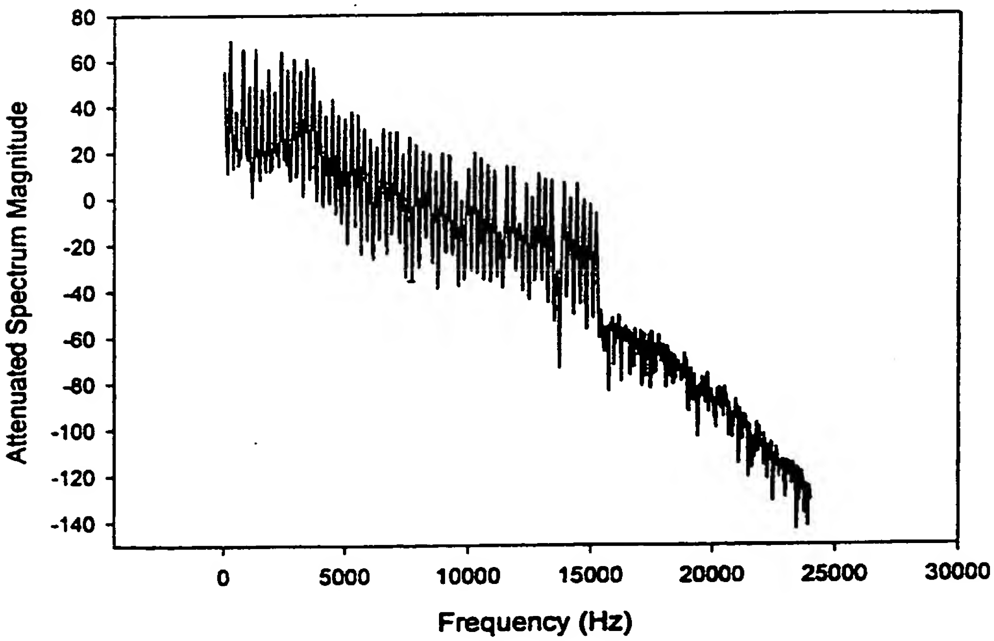


Figure 6. Middle ear attenuation of TestA spectrum

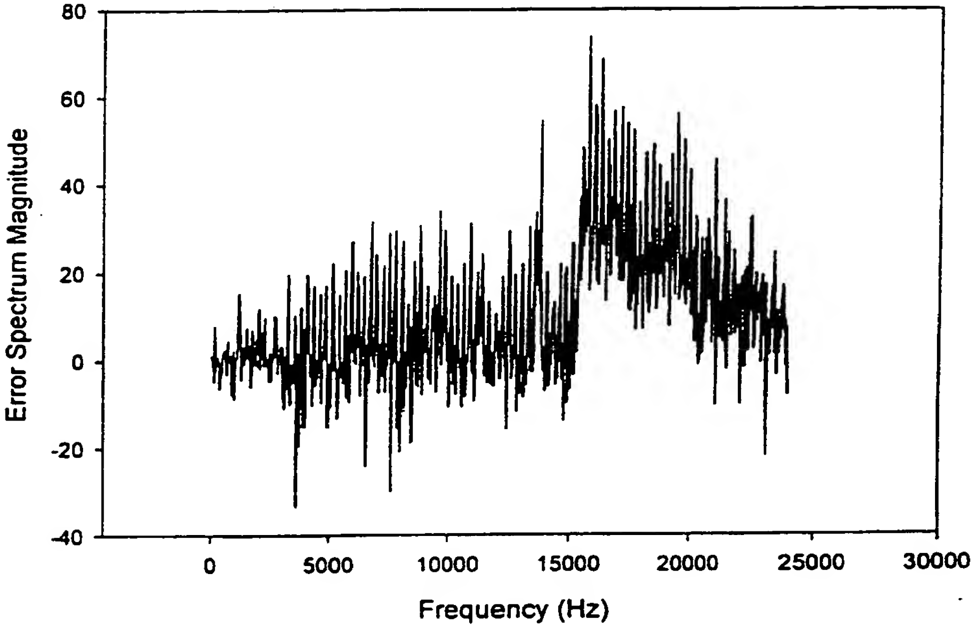


Figure 7. TestA error spectrum

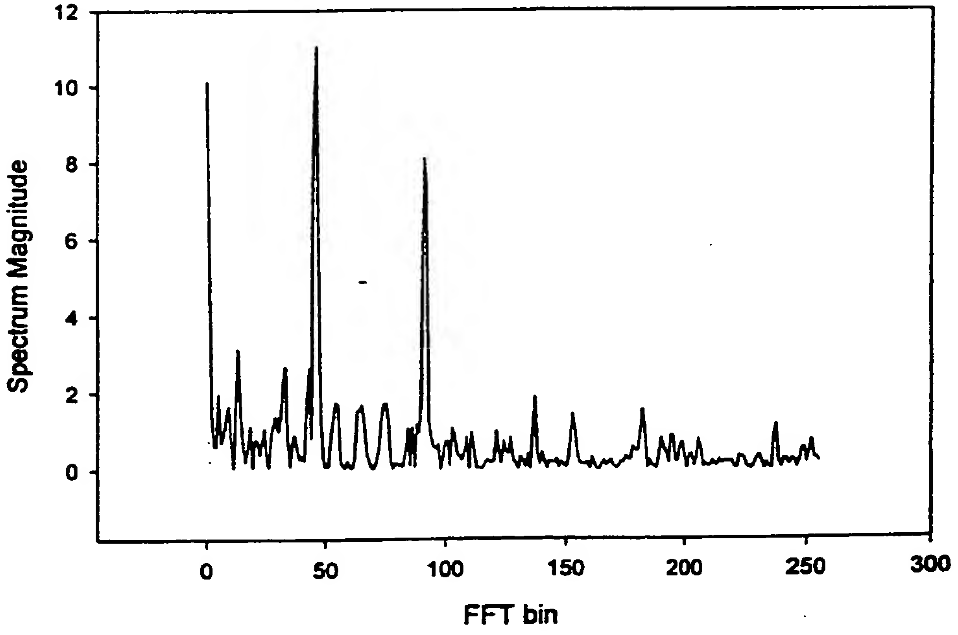


Figure 8. TestA error cepstrum

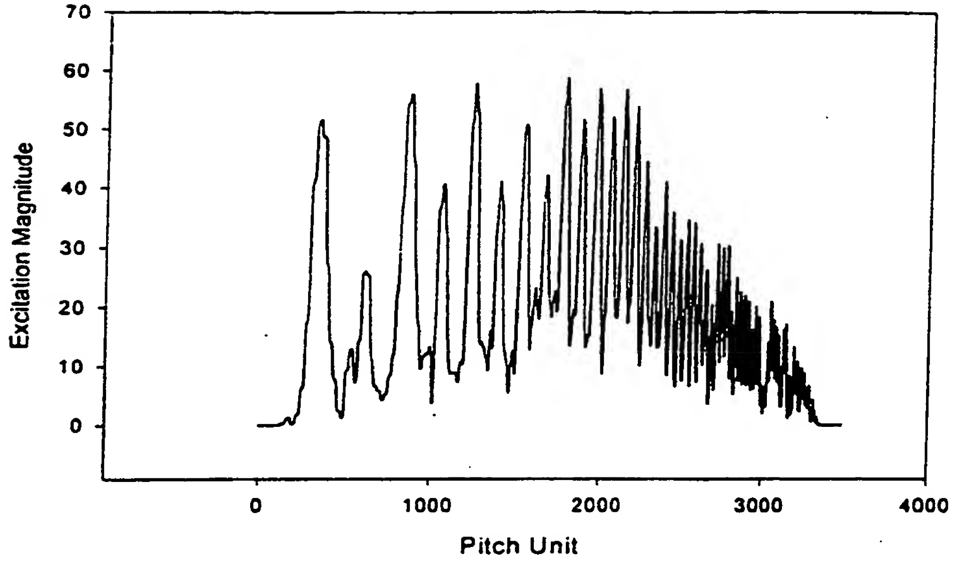


Figure 9. Reference excitation

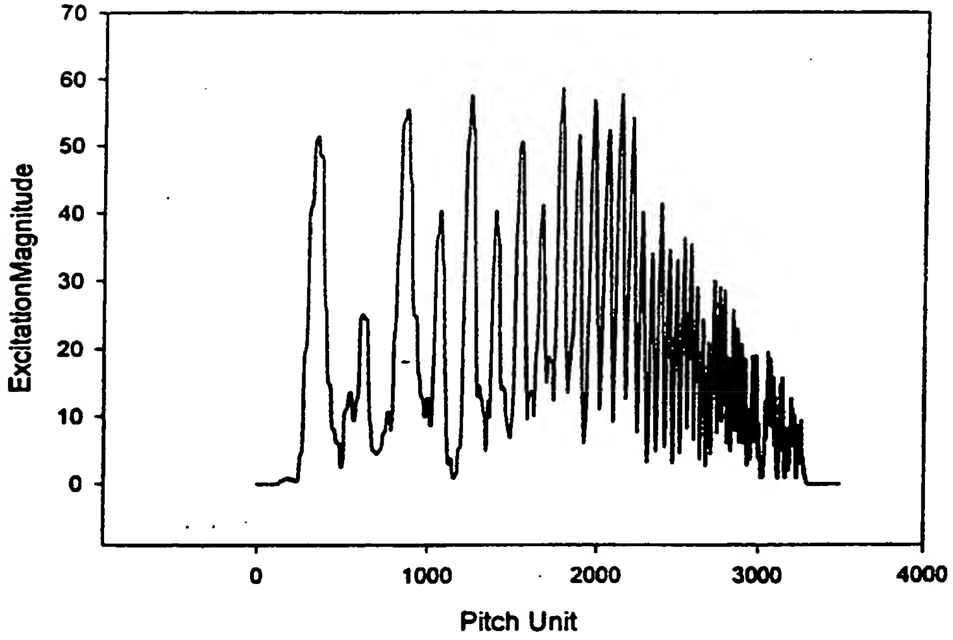


Figure 10. TestA excitation

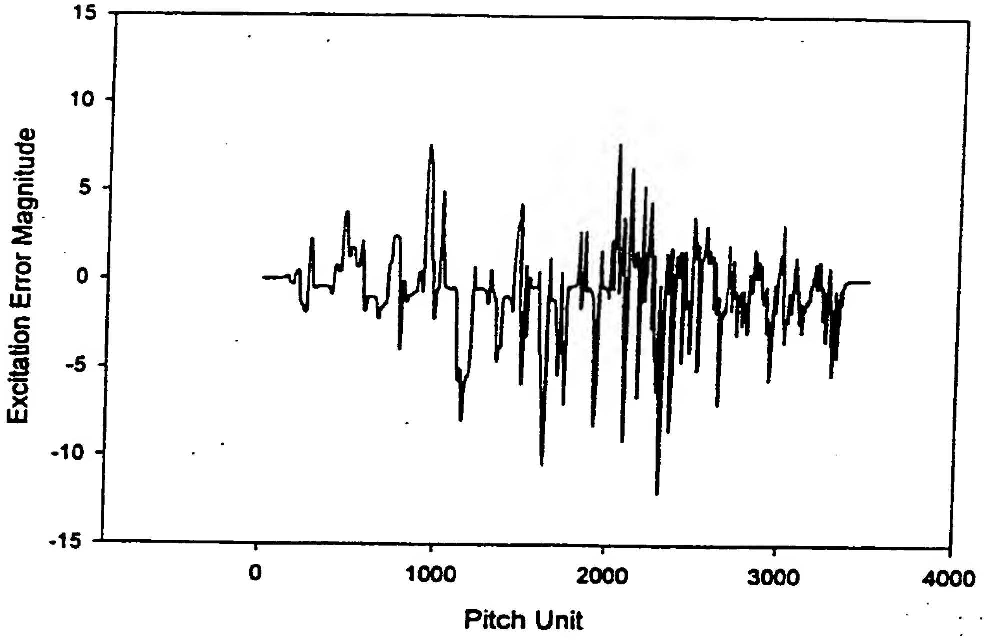


Figure 11. TestA excitation error

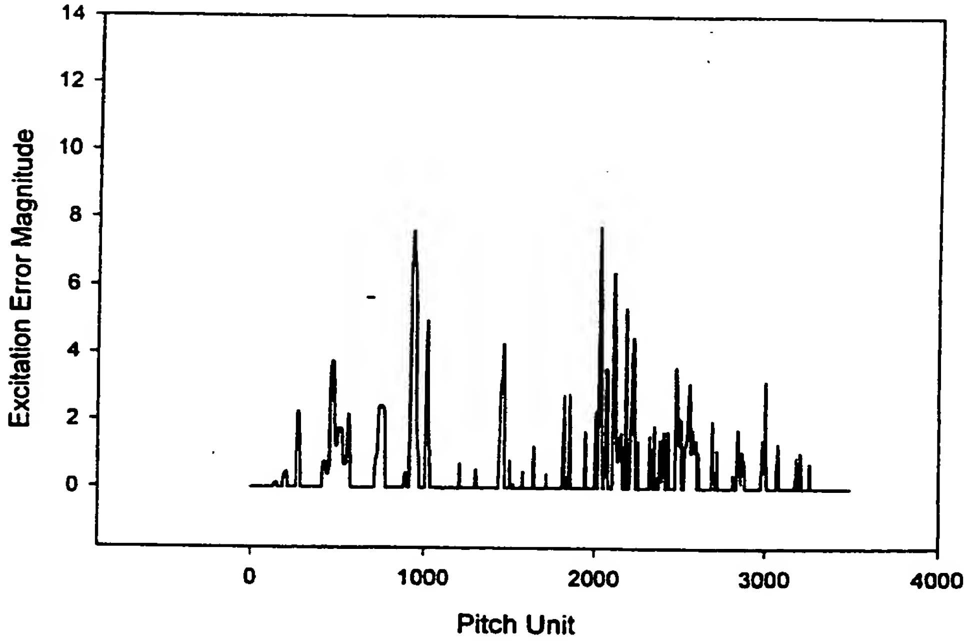


Figure 12. TestA echoic memory output

INTERNATIONAL SEARCH REPORT

Inte onal Application No
PCT/CA 99/00258

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 G10L3/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	BEERENDS J G ET AL: "A PERCEPTUAL AUDIO QUALITY MESURE BASED ON A PSYCHOACOUSTIC SOUND REPRESENTATION" JOURNAL OF THE AUDIO ENGINEERING SOCIETY, vol. 40, no. 12, 1 December 1992, pages 963-978, XP000514954 cited in the application see figure 7 --- -/--	1,7

☒ Further documents are listed in the continuation of box C.

□ Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search 28 June 1999	Date of mailing of the international search report 12/07/1999
Name and mailing address of the ISA European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Authorized officer Krembel, L

INTERNATIONAL SEARCH REPORT

Inter. nal Application No
PCT/CA 99/00258

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	PETERSEN K T ET AL: "Objective speech quality assessment of compounded digital telecommunication systems" 1997 IEEE FIRST WORKSHOP ON MULTIMEDIA SIGNAL PROCESSING (CAT. NO.97TH8256), PROCEEDINGS OF FIRST SIGNAL PROCESSING SOCIETY WORKSHOP ON MULTIMEDIA SIGNAL PROCESSING, PRINCETON, NJ, USA, 23-25 JUNE 1997, pages 137-142, XP002107379 ISBN 0-7803-3780-8, 1997, New York, NY, USA, IEEE, USA see figure 1 * Paragraph "Objective measure for the total quality" * ---	1,7
A	MEKY M M ET AL: "Prediction of speech quality using radial basis functions neural networks" PROCEEDINGS. SECOND IEEE SYMPOSIUM ON COMPUTERS AND COMMUNICATIONS (CAT. NO.97TB100137), PROCEEDINGS SECOND IEEE SYMPOSIUM ON COMPUTER AND COMMUNICATIONS, ALEXANDRIA, EGYPT, 1-3 JULY 1997, pages 174-178, XP002107380 ISBN 0-8186-7852-6, 1997, Los Alamitos, CA, USA, IEEE Comput. Soc, USA -----	1,7